PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 : C12N 15/29, 15/10 — A2

(11) International Publication Number: WO 94/10315

(43) International Publication Date: 11 May 1994 (11.05.94)

(21) International Application Number: PCT/US93/10090

(22) International Filing Date: 22 October 1993 (22.10.93)

(30) Priority data:
07/965,664 23 October 1992 (23.10.92) US

(81) Designated States: AT, AU, BB, BG, BR, BY, CA, CH, CZ, DE, DK, ES, FI, GB, HU, JP, KP, KR, KZ, LK, LU, LV, MG, MN, MW, NL, NO, NZ, PL, PT, RO, RU, SD, SE, SK, UA, UZ, VN, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

(71) Applicant: PIONEER HI-BRED INTERNATIONAL, INC. [US/US]; 700 Capital Square, 400 Locust Street, Des Moines, IA 50309 (US).

(72) Inventor: BALLO, Barbara ; 1444 41st Place, Des Moines, IA 50311 (US).

(74) Agents: YATES, Michael, E. et al.; 700 Capital Square, 400 Locust Street, Des Moines, IA 50309 (US).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: PROCESS FOR ENHANCING THE CONTENT OF A SELECTED AMINO ACID IN A SEED STORAGE PROTEIN

(57) Abstract

Methods which allow for nutritional improvement of plants and plant tissue by increasing the amount of a selected amino acid(s) in a seed storage protein involve altering a naturally-occurring seed storage protein gene. Oligonucleotides coding for the protein are assembled by use of overlapping synthesized DNA sequences.

```
    PstI
TAACTGCAG ATG GCA AAC ATT TCT GTG GTT GCT GCT GCA CTA CTG GTC      48
          Met Ala Asn Ile Ser Val Val Ala Ala Ala Leu Leu Val
           1               5                   10

TTG CTG GTG TTG GGT CAT GCC ACT GCA AGC ATC TAC AGG ACA GTT GTG    96
Leu Leu Val Leu Gly His Ala Thr Ala Ser Ile Tyr Arg Thr Val Val
     15                  20                  25

GAG TTT GAA GAG GAT GAT GCC ACC AAC CCA ATA GGT CCT AAG ATG AGG   144
Glu Phe Glu Glu Asp Asp Ala Thr Asn Pro Ile Gly Pro Lys Met Arg
 30                  35                  40                  45

AAA TGC AGA AAG GAG TTC CAG AAG GAA CAA ATG TTG AGA GCT TGC CAA   192
Lys Cys Arg Lys Glu Phe Gln Lys Glu Gln Met Leu Arg Ala Cys Gln
 50                  55                  60                  65

CAA TGG TTG AGG AAA CAA GCT AGA CAA GGA AGA TCT GAT GAA TTT GAC   240
Gln Trp Leu Arg Lys Gln Ala Arg Gln Gly Arg Ser Asp Glu Phe Asp
 70                  75                  80                  85

TTT GAA GAT GAC ATG GAG AAT CCT CAA GGA CCA CAG CAG AGA CCT CCT   288
Phe Glu Asp Asp Met Glu Asn Pro Gln Gly Pro Gln Gln Arg Pro Pro
                 90                  95                 100

CTC CTT CAG AAG TGC TGT GAG CAA CTC AAA CAG ATG CAA TCT CAG TGT   336
Leu Leu Gln Lys Cys Cys Glu Gln Leu Lys Gln Met Gln Ser Gln Cys
            105                 110                 115

GTT TGC CCA ACC CTT AAA GGT GCC AGC AAA GCT GTG AAA CAG GAA GAG   384
Val Cys Pro Thr Leu Lys Gly Ala Ser Lys Ala Val Lys Gln Glu Glu
        120                 125                 130

CAG CAA CAA GGC CAG CAA CAA GGT AAG CAG CAG ATG GTT AGG AAG ATC   432
Gln Gln Gln Gly Gln Gln Gln Gly Lys Gln Gln Met Val Arg Lys Ile
    135                 140                 145

TAT AAG ACT GCC AAA CAC CTT CCT AAA GTC TGT GAC ATT CCA CAG GTT   480
Tyr Lys Thr Ala Lys His Leu Pro Lys Val Cys Asp Ile Pro Gln Val
150                 155                 160                 165
                                                    EcoRI
GAT GTA TGC CCA TTT CAG AAG ACC ATG CCT GGG CCC TCA TAC TAGAATT   529
Asp Val Cys Pro Phe Gln Lys Thr Met Pro Gly Pro Ser Tyr ***
                170                 175

CAAT                                                              533
```

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 : C12N 15/29, 15/10	A3	(11) International Publication Number: **WO 94/10315** (43) International Publication Date: 11 May 1994 (11.05.94)

(21) International Application Number: PCT/US93/10090

(22) International Filing Date: 22 October 1993 (22.10.93)

(30) Priority data:
07/965,664 23 October 1992 (23.10.92) US

(71) Applicant: PIONEER HI-BRED INTERNATIONAL, INC. [US/US]; 700 Capital Square, 400 Locust Street, Des Moines, IA 50309 (US).

(72) Inventor: BALLO, Barbara ; 1444 41st Place, Des Moines, IA 50311 (US).

(74) Agents: YATES, Michael, E. et al.; 700 Capital Square, 400 Locust Street, Des Moines, IA 50309 (US).

(81) Designated States: AT, AU, BB, BG, BR, BY, CA, CH, CZ, DE, DK, ES, FI, GB, HU, JP, KP, KR, KZ, LK, LU, LV, MG, MN, MW, NL, NO, NZ, PL, PT, RO, RU, SD, SE, SK, UA, UZ, VN, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published

With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(88) Date of publication of the international search report: 15 September 1994 (15.09.94)

(54) Title: PROCESS FOR ENHANCING THE CONTENT OF A SELECTED AMINO ACID IN A SEED STORAGE PROTEIN

(57) Abstract

Methods which allow for nutritional improvement of plants and plant tissue by increasing the amount of a selected amino acid(s) in a seed storage protein involve altering a naturally-occurring seed storage protein gene. Oligonucleotides coding for the protein are assembled by use of overlapping synthesized DNA sequences.

```
 PstI
TAACTGCAG ATG GCA AAC ATT TCT GTG GTT GCT GCT GCA CTA CTG GTC      48
          Met Ala Asn Ile Ser Val Val Ala Ala Ala Leu Leu Val
           1               5                  10

TTG CTG GTG TTG GGT CAT GCC ACT GCA AGC ATC TAC AGG ACA GTT GTG    96
Leu Leu Val Leu Gly His Ala Thr Ala Ser Ile Tyr Arg Thr Val Val
     15                  20                  25

GAG TTT GAA GAG GAT GAT GCC ACC AAC CCA ATA GGT CCT AAG ATG AGG   144
Glu Phe Glu Glu Asp Asp Ala Thr Asn Pro Ile Gly Pro Lys Met Arg
 30                  35                  40                  45

AAA TGC AGA AAG GAG TTC CAG AAG GAA CAA ATG TTG AGA GCT TGC CAA   192
Lys Cys Arg Lys Glu Phe Gln Lys Glu Gln Met Leu Arg Ala Cys Gln
 50                  55                  60                  65

CAA TGG TTG AGG AAA CAA GCT AGA CAA GGA AGA TCT GAT GAA TTT GAC   240
Gln Trp Leu Arg Lys Gln Ala Arg Gln Gly Arg Ser Asp Glu Phe Asp
 70                  75                  80                  85

TTT GAA GAT GAC ATG GAG AAT CCT CAA GGA CCA CAG CAG AGA CCT CCT   288
Phe Glu Asp Asp Met Glu Asn Pro Gln Gly Pro Gln Gln Arg Pro Pro
                 90                  95                 100

CTC CTT CAG AAG TGC TGT GAG CAA CTC AAA CAG ATG CAA TCT CAG TGT   336
Leu Leu Gln Lys Cys Cys Glu Gln Leu Lys Gln Met Gln Ser Gln Cys
            105                 110                 115

GTT TGC CCA ACC CTT AAA GGT GCC AGC AAA GCT GTG AAA CAG GAA GAG   384
Val Cys Pro Thr Leu Lys Gly Ala Ser Lys Ala Val Lys Gln Glu Glu
        120                 125                 130

CAG CAA CAA GGC CAG CAA CAA GGT AAG CAG CAG ATG GTT AGG AAG ATC   432
Gln Gln Gln Gly Gln Gln Gln Gly Lys Gln Gln Met Val Arg Lys Ile
    135                 140                 145

TAT AAG ACT GCC AAA CAC CTT CCT AAA GTC TGT GAC ATT CCA CAG GTT   480
Tyr Lys Thr Ala Lys His Leu Pro Lys Val Cys Asp Ile Pro Gln Val
150                 155                 160                 165
                                                         EcoRI
GAT GTA TGC CCA TTT CAG AAG ACC ATG CCT GGG CCC TCA TAC TAGAATT   529
Asp Val Cys Pro Phe Gln Lys Thr Met Pro Gly Pro Ser Tyr ***
                170                 175

CAAT                                                              533
```

PROCESS FOR ENHANCING THE CONTENT OF A SELECTED AMINO ACID IN A SEED STORAGE PROTEIN

Technical Field

The present invention relates to methods of producing transgenic plants having an increased content of selected amino acids in modified seed storage proteins and, more particularly, to methods of making an improved seed storage protein.

Background of the Invention

Greater recognition of the role of plants in supplying essential amino acids to the animal world has led to emphasis on the development of new food plants that have proteins that are better balanced for human and animal nutrition. Classical plant breeding techniques have limitations for achieving this goal. Molecular genetics, however, shows potential for overcoming these limitations.

Seed storage proteins represent up to 90% of total seed protein in many plant seeds. Shotwell and Larkins (1989) In: The Biochemistry of Plants Vol. 15 (Academic Press, San Diego: Stumpf and Conn, eds.) Chapter 7: 29. These naturally-occurring proteins are used as a source of nutrition for young seedlings for the growth period just following germination. The genes encoding them are strictly regulated, being expressed in a highly tissue-specific and developmentally stage-specific fashion. Walling, et al. (1986) Proc. Natl. Acad. Sci. 83, 2123-2127; Higgins, T.J.V. (1984) Ann. Rev. Plant Physiol. 35, 191-221. Thus they are expressed almost exclusively in developing seed, and different classes of seed storage proteins may be expressed at different stages in the development of the seed.

The expression of foreign genes in plants is well

established. D Blaere, et al. (1987) Methods in Enzymology 153, 277. Seed storage protein genes have been transferred to other plants. Okamura, et al. (1986) Proc. Natl. Acad. Sci. 83, 8240; Sengupta-Gopalan, et al. (1985) Proc. Natl. Acad. Sci. 82, 3320; Higgins, et al. (1988) Plant Mol. Biol. 11, 683; Ellis, et al. (1988) Plant Mol. Biol. 10, 203; Barker, et al. (1988) Proc. Natl. Acad. Sci. 85, 458; Vandekerckhove, et al. (1989) Bio/Technol. 7, 929; and Altenbach, et al. (1989) Plant Mol. Biol. 13, 513. In most of these cases it was shown that within its new environment, the transferred seed storage protein gene is expressed in a tissue-specific and developmentally regulated manner. Beachy, et al. (1985) EMBO J. 4, 3047. The expression levels varied, but reached as high as 8% of the total seed protein. Altenbach, et al., supra; Voelker, et al. (1989) Plant Cell 1, 95.

However, design of a synthetic seed storage protein requires more than mere substitution of the desired amino acid for naturally-occurring amino acids in the target protein. Criteria must be defined for maximizing the potential of success and the ultimate expression of the gene in the targeted host plant. Even selection of the class of storage proteins least likely to present difficulties is important, and is dependent on the availability of sequence data for that class of proteins, the relative gene size within that class, and the degree of processing and post-translational modification necessary for deposition. Seed storage proteins are nominally classified by density gradient sedimentation values: 2S, 7S, and 11S. Although the 7S and 11S proteins tend to be one general type per sedimentation value, the 2S seed proteins are a diverse group. The 2S sedimentation value implies a relatively low molecular weight, and the 2S proteins include classic storage proteins as well as lectins, protease inhibitors, and others. The 2S storage proteins appear to be less restricted in amino acid composition than 7S and 11S proteins, and include species which are relatively rich in basic amino acids. Additionally, the 2S storage proteins

are encoded on small genes, making the prospect of synthesizing a new 2S gene from oligonucleotides attractive.

Among published seed protein sequence data, no protein incorporating a non-limiting amount of lysine has been identifi d. Lysine comprises from 3 to 7% of the total amino acids in known seed protein sequences. It is estimated that a protein containing 10 to 15% lysine, expressed transgenically at a level of 2 to 5%, is necessary to cause a noticeable increase in seed deposition of lysine. No storage protein-coding sequence which meets this criterion is known.

Storage proteins can be modified by incorporating inserts containing one or more selected amino acids such as lysine, resulting in a lysine-rich polypeptide that can be transferred into plant cells. Or, following the design of a storage protein with a known sequence, a lysine-rich polypeptide can be synthesized by substitution of specific amino acids and transferred into a host cell.

There is a recognized need for lysine-rich seed storage proteins and for an efficient, accurate method of producing the same. Further, there is also a recognized need for a method to produce a DNA or cDNA sequence that codes for an increased amount of any essential amino acid that can be expressed transgenically as a seed storage protein. A DNA "coding sequence" is a DNA sequence which is transcribed and translated into a polypeptide in vivo when placed under the control of appropriate regulatory sequences. The boundaries of the coding sequence are determined by the ATG start codon at the 3' terminus. Examples of coding sequences include cDNA, genomic DNA sequences from cells, and synthetic DNA sequences.

When designing sequences to be rich in certain amino acids, care must be taken that the substitutions with the selected amino acids does not influence the stability of the modified 2S protein. Certain insertions, such as, long stretches of particular amino acids, may result in shapes and turns which would cause instability, poor expression, or poor accumulation due to disruption in normal folding

patterns of the protein. In addition, replacement must be conservative in that hydrophobic amino acids and those giving charge and polarity are not substituted so that the overall structure and stability of the molecule will not be adversely affected. Polarity and direction are due to acidic (negative) and basic (positive) charges on the amino acid residues.

To synthesize DNA molecules, the two complementary strands are constructed separately because only single-stranded DNA (oligonucleotides) can be synthesized. These are then hybridized (by formation of hydrogen bonds) and linked to larger DNA units by enzymatic coupling in order to construct genes or their regulatory units. A gene is a DNA sequence responsible for the production of polypeptides. It is now possible, given the various DNA recombination techniques, to construct any given gene, whether synthetic or natural, to reproduce it, and to convert it into polypeptides using whole cell systems.

Oligonucleotides are polymers built up by the polycondensation of nucleoside phosphates. In the past, the majority of synthetic genes have been assembled using complementary oligonucleotides which represent both entire strands. Gapped fill-ins refer to the pairing of complementary nucleotides along sections of DNA where pairing is incomplete (single-stranded sections) to form complementary DNA strands for those segments. Gapped fill-ins have been published only for single pairs of overlapping oligonucleotides, which limits the length of the target molecule. Thus, construction of long synthetic sequences required subcloning (moving a sequence from one vector to another to produce copies) and/or pasting together of regions via restriction sites. The only method utilizing an overlap extension procedure requires splicing of double-stranded gene fragments. Horton, et al. (1989) Gene 77, 61.

The sequential extension method presented by this invention obviates the subcloning requirement, and allows simple, one day assembly of larger gene regions. This method is approximately 30% more cost-effective even without

consideration of personnel time than the usual method of assembling complete complementary oligonucleotides because it allows enzymatic synthesis of gap regions. Khorana (1968) Pure Appl. Chem. 17, 349. A more recent publication offers similar cost savings by incorporation of a terminal 3′ hairpin structure to prime synthesis of the second strand. However that method is limited by the length of oligonucleotides. Ulhmann, et al. (1988) Gene 71, 29. Another method utilizes short overlap regions to prime polymerase, but both of these methods rely on ligation of separate double-stranded regions for assembly. Rink, et al. (1984) Nucleic Acid Res. 12, 16; Rossi, et al. (1982) J. Biol. Chem. 257, 9226. A third method relies on *in vivo* gap repair, and requires that one strand of synthetic DNA be complete, though it may contain nicks bridged by short oligonucleotides of the opposite strand. It has only been used to assemble a 270 bp fragment. Adams, et al. (1989) Nucleic Acid Res. 16, 4287.

The advantages of this invention are: (a) it is cost effective because fewer oligonucleotides are required and less time is spent in oligonucleotide preparation because crude oligonucleotides work well; (b) it is a simple two-reaction (extension/amplification) procedure that is complete in 1-2 days; (c) it does not require that restriction sites for assembly by ligation be included in gene design, hence no unnecessary mutations are introduced; (d) it enables rapid inclusion of degenerate oligonucleotide regions if desired, without separate assembly or cloning reactions; and (e) it enables the assembly of chimeric genes without the introduction of mutagenic restriction sites, i.e., it enables "perfect" promoter-gene fusions.

The present invention further provides improvements in the nutritional value of edible organisms, including, but not limited to, higher plants. In particular, the present invention provides for the assembly of synthetic oligonucleotides by means of overlapping sequences, including the nucleic acid sequences encoding the lysine-rich proteins.

In one embodiment, the present invention provides

nucleic acids in the form of a DNA molecule, which encode one or more subunits of a lysine-rich (approximately 14%) 2S seed storage-type protein. Other isoforms will be at least about 80% homologous at the amino acid sequence level to this representative member, preferably at least about 85% homologous, and more preferably at least about 90% homologous.

In a further embodiment, the present invention provides a cell comprising a replicon containing the chemically-synthesized, lysine-rich 2S storage protein combined with a promoter which includes regulatory sequences that provide for the expression of said protein in said cell, said subunit being heterologous to said cell. In particularly preferred embodiments, the cellular host is a higher plant or animal cell.

Brief Description of the Figures

Figure 1 shows the complete nucleic acid sequence of a 2S seed storage protein with increased lysine content. The double-stranded molecule is cleaved with restriction enzymes (PstI and EcoRI) at bases as indicated to allow cloning.

Amino acid residues are numbered beneath the sequence. Mature protein is comprised of residues 39-74 (small subunit) crosslinked via S-S bonds to residues 85-170 (large subunit). Residues 1-38 constitute a signal sequence and N-terminus processed site. Residues 75-84 constitute a "linker" type peptide, which is excised as the protein folds. Residue 171 is a carboxy-terminal residue which is also excised at protein maturity.

Figure 2 shows the oligonucleotides used in construction of the 2S seed protein and their design as pairs sharing complementary overlap regions of 17-24 nucleotides. Each pair has a similar overlap with the adjacent pair.

Figure 3 shows the first and second sequential extension products that are formed as the six extensions are implemented.

Figure 4 shows the third, fourth, fifth, and sixth

sequential extension products that are formed as the six extensions are completed.

Disclosure of the Invention

In addition to the techniques described below, the practice of the present invention will employ conventional techniques of molecular biology, microbiology, recombinant DNA technology, and plant science, all of which is within the skill of the art. Such techniques are explained fully in the literature. See, e.g., Maniatis et al., Molecular Cloning: A Laboratory Manual (1982); DNA Cloning: Volume I and II (D.N. Glover, ed., 1985); Oligonucleotide Synthesis (M.J. Gait, ed., 1984): Nucleic Acid Hybridization (B.D. Hames & S.J. Higgins, eds., 1985); Transcription and Translation (B.D. Hames & S.J. Higgins, eds, 1984): Animal Cell Culture (R.I. Freshney, ed., 1986); Plant Cell Culture (R.A. Dixon, ed., 1985); Propagation of Higher Plants Through Tissue Culture (K.W. Hughes et al., eds., 1978); Cell Culture and Somatic Cell Genetics of Plants (I.K. Vasil, ed., 1984); Fraley et al. (1986) CRC Critical Reviews in Plant Sciences 4, 1; Biotechnology in Agricultural Chemistry: ACS Symposium Series 334 (LeBaron et al. eds. 1987) the disclosures of which are well-known and are hereby incorporated herein by reference.

The design of the prototypical synthetic plant gene herein is based on published regulatory sequences, including reported enhancer (repetitive) regions found uniquely in seed storage genes. In addition, computer modeling of both hydropathy and evolutionary relatedness of known seed proteins was used in the planning of potential coding sequences, as well as inclusion of codon biases found in published storage protein gene sequences.

With respect to the choice of the regions to be modified, the present invention varies significantly from other work which has been done in this field. European Patent Application No. 318, 341 describes a method for replacement or supplementation of the hypervariable region of a 2S

albumin gene. Based on a model of the Arabidopsis thaliana 2S albumin, the hypervariable region is defined as a section of the large subunit of the protein between the sixth and seventh cysteine residues where little conservation of amino acids is observed. A non-conserved region is a region wherein the nucleotide sequence can be modified either by insertion into it or replacement of a nucleotide sequence which, at least in part, may be foreign to the natural nucleic acid encoding the precursor of the 2S albumins of the plant cells concerned and encodes the appropriate amino acids, without disturbing the stability and correct processing of the storage protein or its transport into parts of the cell. The modification procedure is called site-directed mutagenesis.

The synthetic gene sequence was constructed by the general process of sequential extension of overlapping 3′ ends using DNA polymerase. The sequence was designed to be assembled from six pairs of synthetic oligonucleotides (partial sequences), each having 3′ overlap within the pair, as well as 3′ overlap between adjacent pairs. Assembly is comprised of three parts: filling in pairs to create double-stranded segments; combining all duplexed segments and sequentially extending to form a small number of full length genes; and amplifying (PCR) complete molecules to a quantity sufficient for cloning. It is an efficient and streamlined procedure, useful for constructing large genes with little or no possibility of misjoinder and without the need for intermediate vectors. Numerous pairs of partial sequences can be used to assemble large synthetic genes. There is no limit to the size of predetermined gene structure that this synthetic strategy will allow. Accordingly, it is anticipated that this invention will find important utilization by those skilled in the art.

In one embodiment, each pair is filled in by combining two (paired) oligonucleotides (100 pmol each) in a suitable solution for bonding, comprising 15 μM each dNTP, 40mM Tris-Cl pH7.5, 20mM $MgCl_2$, 50 mM NaCl, and 10mM DTT (25 μl final volume). The oligonucleotide mix is heat denatured (95°C)

and allowed to anneal by slowly cooling to room temperature. Heat-sensitive DNA polymerase (examples: _E. coli_ "Klenow", Sequenase {Registered: US Biochemical}) is added (1.5 U) and the reaction allowed to proceed 10 minutes at room temperature. Alternatively, heat-stable polymerase (e.g., Taq polymerase) may be substituted if the buffer is replaced with 50mM KCl, 10mM Tris-Cl pH8.3, 1.5mM $MgCl_2$, .01% BSA, and the reaction mix annealed at 55°C and extended at 72°C. Sequential extension of these pairs is accomplished by combining aliquots of each of the above reactions, adding sufficient dNTPs, and sequentially heating, reannealing, and extending in the presence of polymerase. This is easily accomplished using Taq polymerase and commercially available heat cycling blocks (e.g., DNA Thermal Cycler {Perkin-Elmer/Cetus}), and requires buffer adjustment as noted above. Heat-labile polymerase may be substituted, but requires manual transfer of tubes between heat blocks of suitable temperature. The number of cycles required to generate full length sequences is dependent on the number of duplexed components, and is minimally half that number. To generate sufficient full length molecules to allow gel detection, the molecules must be cycled a greater number of times. In the example from the previous paragraph, the partial sequences were sequentially extended for a total of 12 cycles in order to discern full length molecules. Obtaining a clonable amount of this gene sequence is possible using PCR, and requires only a small portion (2%) of the sequential extension reaction as template.

Modes for carrying out the Invention

Example 1

Design of the protein

A putative 2S seed storage protein sequence was derived from published protein sequences, Crouch, et al. (1983) _J._

Mol. Appl. Gen. 2, 273; Ericson, et al. (1986) J. Biol. Chem. 261, 14576; Altenbach, et al. (1987) Plant Mol. Biol. 8, 239; Krebbers, et al. (1988) Plant Physiol. 87, 859), and by using peptide sequence data from various *Brassica* spp. obtained in this laboratory (unpublished). These members of the 2S class of seed storage proteins are synthesized as precursor polypeptides of 15-21 kDa and undergo a number of processing steps to yield the stored protein, comprised of a large and a small subunit of combined MW of 9-17 kDa. The proposed protein sequence (Figure 1) includes all processing regions typical of such 2S seed proteins. The first 22 amino acids should function as a transit peptide to direct protein inclusion in storage bodies (Chrispeels, et al. 1982 J. Cell Biol. 93:306). In addition to the first 22 amino acids, residues 23-38, 75-84, and 171 are those amino acids which should be deleted in the final stored product by processing steps typical of these 2S seed proteins. The accumulated protein should thus be two subunits of 4.4 kDa (residues 39-74) and 9.7 kDa (residues 85-170). Codons were selected for the synthetic gene based on observed codon biases in seed storage proteins (data not shown).

Example 2

Synthesis of oligonucleotides

Oligonucleotides from 56 to 69 nucleotides in length were synthesized on an Applied Biosystems Model 380B synthesizer, deblocked, treated with ammonia at 50°C, vacuum-dried and resuspended in water. The oligonucleotides were used with no further purification.

Oligonucleotides used in this construction were designed as pairs sharing complementary overlap regions of 17-24 nucleotides, each pair having a similar overlap with the adjacent pair (Figure 2). Following denaturation and annealing with all oligomers present in the reaction, molecules of the most stable duplex structure formed, and

allowed extension of the duplex from the overlaps. Repetition of such extensions produced successively longer molecules, hence progressively larger regions of complementation. Sequential extension products are shown schematically in Figure 3. The first extension reaction can yield only those products shown, and required polymerase fill-ins of 37-51 nucleotides from overlap regions of 17-24 base pairs in the claimed synthetic gene. The second round of extension must also proceed from minimal overlaps (17-18 base pairs), with the addition of 79-102 nucleotides to the complementary regions. Beginning with the third extension, progressively larger overlaps were available. Only the longest, hence most stable duplex conformations, are shown in Figure 3. At the end of the third extension reaction some completed molecules were present in the reaction. A total of six extensions increased the probability of obtaining complete sequences.

Example 3

Amplification

An aliquot of the extension products served as a template for in vitro amplification using distal 5′ and 3′ oligonucleotides (oligos 1+ and 6−) as primers. Both the Taq polymerase and the T7 DNA polymerase extension reactions yielded single Taq amplification products of the expected 530 bp.

Example 4

Cloning and expression

The amplification products of Example 3 were gel purified, cut at the PstI and EcoRI sites included at the 5′ and 3′ ends of the synthetic sequence, and cloned into similarly digested pTZ18u. Recombinant plasmids were transfected into DH5α and plated on selective media containing x-Gal. White

colonies were selected for mini-pr ps of DNA, and screened for the pres nce of the 206 bp Bgl2 fragment. Six of the Taq-extended clones and seven of the T7-extended clones were sequenced completely at least once in each direction, and the sequence analysis results are shown in Table 1. One of six clones from Taq extension and one of seven clones from T7 extension contained perfect constructs. The clones from the Taq extension contained a total of 10 induced single base pair mutations: 6 substitutions, 3 deletions and one insertion. The sum mutation rate with Taq extension was thus 10/(6x530) or 1 mutation per 318 nucleotides. T7 extensions contained considerably more mutations, including 10 substitutions, one insertion and 3 deletions of 2, 3 and 9 base pairs. The sum mutation rate with T7 polymerase extensions was thus 25/(7x530) or 1 mutation per 148 nucleotides.

Mini plasmid preps used to screen for the Bgl2 fragment were digested with EcoRI and PstI, Southern blotted and examined by hybridization to a probe prepared from the complete insert of the correct synthetic gene clone, pTL315. It was found that of clones produced by Taq extension, only those possessing the BglII fragment contained any portion of the synthetic gene. However, 24 clones obtained through T7 extension contained some portion of the synthetic gene, and only six of these included the predicted BglII fragment. More amplification products result from the T7 extension mixes than from those of Taq. It is likely that the lower temperature (37°) used for T7 extensions allowed more mismatches during annealing and extension than that allowed during the Taq (72°) extensions.

Table 1

Clones selected from sequential extensions

Designation	Enzyme used in extensions	Mutation	Location[a]
pTL310	Taq	A->T	OL(S)
pTL314	Taq	G deletion	Fl
pTL315	Taq	none	
pTL332	Taq	A deletion	OL(S)
pTL333	Taq	A->G	OL(F)
pTL340	Taq	A deletion	Fl
		C insertion	Fl
pTL344	Taq	G->T	Fl
		A->C	Fl
		CA->AC	Fl
pTl410	Sequenase	CA deletion	OL(S)
		T->C	OL(S)
pTL414	Sequenase	T-<G	Fl
		A insertion	Fl
pTL423	Sequenase	C->T	Fl
		T-<G	OL(F)
pTL459	Sequenase	none	
pTL478	Sequenase	GTG deletion	Fl
		T->C	Fl
pTL652	Sequenase	A->G	Fl
		G->A	Fl
pTL657	Sequenase	9 bp deletion	Fl
		A->G	OL(S)
		T->C	Fl
		C->A	Fl

[a] OL(S):overlap during first sequential extension
OL(F):overlap during paired oligonucleotide fill-in
Fl:fill-in region during either of the above reactions

Industrial Applicability

Directly or indirectly, animals obtain their essential amino acids (those they are unable to synthesize) from eating plants. Most seeds, the major plant protein sources, are deficient in one or more amino acids essential for proper nutrition of higher animals. Dicotyledonous seeds, such as legumes, generally lack sufficient sulfur-containing amino acids (cysteine and methionine), while monocotyledonous plants (cereals) typically lack adequate lysine, as well as tryptophan and threonine. Plants can serve as adequate amino acid sources if complementary seeds (e.g., rice and beans) are ingested simultaneously, and in the proper quantity.

Cereals and legumes are combined in this complementary way in the formulation of diets for swine. Current feeding practices in the United States utilize 85% corn and 15% soybean meal in swine diets. The predominance of corn as the major dietary component is due mainly to its low cost and high carbohydrate content. The low protein levels are supplemented with soybean meal to provide adequate protein nutrition. Because corn is particularly deficient in lysine (2%), added soybean, although sufficient in lysine (6.4%) when used as the sole protein source, cannot raise lysine levels to those necessary for maximum swine growth. Thus swine feed is frequently supplemented with "synthetic" lysine. Current levels of supplemental lysine average about 1kg per metric ton of feed at a cost of $4.50/kg lysine. The U.S. market for lysine (primarily used in feeds) is 20Mkg, resulting in retail sales of $100M. Strategies to reduce this supplementation of lysine include the use of newly developed high-lysine (3.3%) corn varieties. These varieties may obviate the need for lysine addition to feed in the future. However high-lysine varieties have not yet been widely accepted by farmers, because they typically show poor growth and low yield characteristics. Additionally, existing high-lysine corn lines are the result of a recessive mutation, which increases the difficulty of

breeding this characteristic into popular varieties. Therefore, these varieties of corn are an expensive source of high-lysine protein.

A reasonable alternative is to enhance lysine levels in corn, soybean, and other crops through introduction of new seed storage protein genes. For example, soymeal is a component of animal feeds because of its high protein quality and content. A modest increase in soy protein lysine levels may be of great benefit to the feed market due to the high quality protein background in soybean. Molecular biology now provides the tools to alter amino acid composition via gene transfer and provide, through this invention, for the nutritional enhancement of soybeans and other crops.

Sequence Listing

(1) GENERAL INFORMATION:

(i) APPLICANT: Barbara Ballo

(ii) TITLE OF INVENTION: Process for Enhancing the Content of a Selected Amino Acid in a Seed Storage Protein

(iii) NUMBER OF SEQUENCES: 13

(iv) CORRESPONDENCE ADDRESS:

(A) ADDRESSEE: Pioneer Hi-Bred International, Inc.

(B) STREET: 700 Capital Square 400 Locust Street

(C) CITY: Des Moines

(D) STATE: Iowa

(E) COUNTRY: United States

(F) ZIP: 50309

(v) COMPUTER READABLE FORM:

(A) MEDIUM TYPE: Diskette -- 3.5 inch, 720 kb storage

(B) COMPUTER: IBM Compatible

(C) OPERATING SYSTEM: MS-DOS

(D) SOFTWARE: WORDPERFECT

(vi) CURRENT APPLICATION DATE:

(A) APPLICATION NO.

(B) FILING DATE:

(C) CLASSIFICATION:

(viii) ATTORNEY/AGENT INFORMATION:

(A) NAME: Pearlmutter, Nina L.

(B) REGISTRATION NUMBER: 35,639

(C) REFERENCE/DOCKET NUMBER: 0215 US

(ix) TELECOMMUNICATION INFORMATION:

(A) TELEPHONE: (515) 245-3596

(B) TELEFAX: (515) 245-3634

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 533 bases

(B) TYPE: nucleotide

(C) STRANDEDNESS: double

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTI-SENSE: N/A

(xi) SEQUENCE DESCRIPTION:Seq. ID. No. 1

TAACTGCAG ATG GCA AAC ATT TCT GTG GTT GCT GCT GCA CTA CTG GTC 48
Met Ala Asn Ile Ser Val Val Ala Ala Ala Leu Leu Val
1 5 10

TTG CTG GTG TTG GGT CAT GCC ACT GCA AGC ATC TAC AGG ACA GTT GTG 96
Leu Leu Val Leu Gly His Ala Thr Ala Ser Ile Tyr Arg Thr Val Val
15 20 25

GAG TTT GAA GAG GAT GAT GCC ACC AAC CCA ATA GGT CCT AAG ATG AGG 144
Glu Phe Glu Glu Asp Asp Ala Thr Asn Pro Ile Gly Pro Lys Met Arg
30 35 40 45

AAA TGC AGA AAG GAG TTC CAG AAG GAA CAA ATG TTG AGA GCT TGC CAA 192
Lys Cys Arg Lys Glu Phe Gln Lys Glu Gln Met Leu Arg Ala Cys Gln
50 55 60 65

CAA TGG TTG AGG AAA CAA GCT AGA CAA GGA AGA TCT GAT GAA TTT GAC 240
Gln Trp Leu Arg Lys Gln Ala Arg Gln Gly Arg Ser Asp Glu Phe Asp
70 75 80 85

TTT GAA GAT GAC ATG GAG AAT CCT CAA GGA CCA CAG CAG AGA CCT CCT 288
Phe Glu Asp Asp Met Glu Asn Pro Gln Gly Pro Gln Gln Arg Pro Pro
90 95 100

CTC CTT CAG AAG TGC TGT GAG CAA CTC AAA CAG ATG CAA TCT CAG TGT 336
Leu Leu Gln Lys Cys Cys Glu Gln Leu Lys Gln Met Gln Ser Gln Cys
105 110 115

GTT TGC CCA ACC CTT AAA GGT GCC AGC AAA GCT GTG AAA CAG GAA GAG 384
Val Cys Pro Thr Leu Lys Gly Ala Ser Lys Ala Val Lys Gln Glu Glu
120 125 130

CAG CAA CAA GGC CAG CAA CAA GGT AAG CAG CAG ATG GTT AGG AAG ATC 432
Gln Gln Gln Gly Gln Gln Gln Gly Lys Gln Gln Met Val Arg Lys Ile
135 140 145

TAT AAG ACT GCC AAA CAC CTT CCT AAA GTC TGT GAC ATT CCA CAG GTT 480
Tyr Lys Thr Ala Lys His Leu Pro Lys Val Cys Asp Ile Pro Gln Val
150 155 160 165

GAT GTA TGC CCA TTT CAG AAG ACC ATG CCT GGG CCC TCA TAC TAGAATT 529
Asp Val Cys Pro Phe Gln Lys Thr Met Pro Gly Pro Ser Tyr ***
170 175

CAAT 533

(3) INFORMATION FOR SEQ ID NO: 2:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 69 bases

(B) TYPE: nucleotide

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTI-SENSE: No

(xi) SEQUENCE DESCRIPTION:

Seq. ID. No. 2

TAACTGCAGA TGGCAAACAT TTCTCTGGTT GCTGCTGCAC TACTGGTCTT GCTGGTGTTG 60

GGTCATGCC 69

(4) INFORMATION FOR SEQ ID NO: 3:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 69 bases

(B) TYPE: nucleotide

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTI-SENSE: Yes

(xi) SEQUENCE DESCRIPTION:

Seq. ID. No. 3

GGTGGCATCA TCCTCTTCAA ACTCCACAAC TGTCCTGTAG ATGCTTGCAG TGGCATGACC 60

CAACACCAG 69

(5) INFORMATION FOR SEQ ID NO: 4:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 57 bases

(B) TYPE: nucleotide

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTI-SENSE: No

(xi) SEQUENCE DESCRIPTION:

Seq. ID. No. 4

GAAGAGGATG ATGCCACCAA CCCAATAGGT CCTAAGATGA GGAAATGCAG AAAGGAG 57

(6) INFORMATION FOR SEQ ID NO: 5:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 56 bases

(B) TYPE: nucleotide

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTI-SENSE: Yes

(xi) SEQUENCE DESCRIPTION:

Seq. ID. No. 5

CCATTGTTGG CAAGCTCTCA ACATTTGTTC CTTCTGGAAC TCCTTTCTGC ATTTCC 56

(7) INFORMATION FOR SEQ ID NO: 6:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 59 bases

(B) TYPE: nucleotide

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTI-SENSE: No

(xi) SEQUENCE DESCRIPTION:

Seq. ID. No. 6

GAGCTTGCCA ACAATGGTTG AGGAAACAAG CTAGACAAGG AAGATCTGAT GAATTTGAC 59

(8) INFORMATION FOR SEQ ID NO: 7:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 61 bases

(B) TYPE: nucleotide

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTI-SENSE: Yes

(xi) SEQUENCE DESCRIPTION:

Seq. ID. No. 7

GGTCTCTGCT GTGGTCCTTG AGGATTCTCC ATGTCATCTT CAAAGTCAAA TTCATCAGAT 60

C 61

(9) INFORMATION FOR SEQ ID NO: 8:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 57 bases

(B) TYPE: nucleotide

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTI-SENSE: No

(xi) SEQUENCE DESCRIPTION:

Seq. ID. No. 8

GGACCACAGC AGAGACCTCC TCTCCTTCAG AAGTGCTGTG AGCAACTCAA ACAGATG 57

(10) INFORMATION FOR SEQ ID NO: 9:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 64 bases

(B) TYPE: nucleotide

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTI-SENSE: Yes

(xi) SEQUENCE DESCRIPTION:

Seq. ID. No. 9

CAGCTTTGCT GGCACCTTTA AGGGTTGGGC AAACACACTG AGATTGCATC TGTTTGAGTT 60

GCTC 64

(11) INFORMATION FOR SEQ ID NO: 10:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 65 bases

(B) TYPE: nucleotide

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTI-SENSE: No

(xi) SEQUENCE DESCRIPTION:

Seq. ID. No. 10

AAGGTGCCAG CAAAGCTGTG AAACAGGAAG AGCAGCAACA AGGCCAGCAA CAAGGTAAGC 60

AGCAG 65

(12) INFORMATION FOR SEQ ID NO: 11:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 56 bases

(B) TYPE: nucleotide

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTI-SENSE: Yes

(xi) SEQUENCE DESCRIPTION:

Seq. ID. No. 11

GGAAGGTGTT TGGCAGTCTT ATAGATCTTC CTAACCATCT GCTGCTTACC TTGTTG 56

(13) INFORMATION FOR SEQ ID NO: 12:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 59 bases

(B) TYPE: nucleotide

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTI-SENSE: No

(xi) SEQUENCE DESCRIPTION:

Seq. ID. No. 12

GACTGCCAAA CACCTTCCTA AAGTCTGTGA CATTCCACAG GTTGATGTAT GCCCATTTC 59

(14) INFORMATION FOR SEQ ID NO: 13:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 57 bases

(B) TYPE: nucleotide

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: synthetic DNA

(iii) HYPOTHETICAL: No

(iv) ANTI-SENSE: Yes

(xi) SEQUENCE DESCRIPTION:

Seq. ID. No. 13

ATTGAATTCT AGTATGAGGG CCCAGGCATG GTCTTCTGAA ATGGGCATAC ATCAACC 57

WHAT IS CLAIMED IS:

1. A method of making an improved seed storage protein by altering a naturally-occurring seed storage protein having a known amino acid sequence to increase its content of a selected amino acid, comprising the steps of:

a. identifying conserved, non-conserved and hypervariable residues in the amino acid sequence of the naturally-occurring protein by comparison of the amino acid sequence of the protein with amino acid sequences of other homologous seed storage proteins; and

b. replacing one or more non-conserved DNA residues coding for the protein with DNA residues coding for the selected amino acid, provided that

i) the replacement is conservative with respect to hydrophobicity, polarity and charge, and

ii) the replacement does not create any pairs of adjacent amino acids which are not found in the naturally-occurring seed storage protein or the homologous seed storage proteins.

2. A method according to claim 1 comprising the further steps of synthesizing a DNA sequence which codes for the altered seed storage protein and synthesizing the altered seed storage protein by transcription and translation of the DNA sequence in a living cell.

3. A method according to claim 2 wherein the DNA sequence is synthesized by site-directed mutagenesis of a DNA sequence which codes for the naturally-occurring seed storage protein.

4. A method according to Claim 2 wherein the DNA sequence which codes for the naturally-occurring seed storage protein is genomic DNA.

5. A method according to Claim 3 wherein the DNA sequence which codes for the naturally-occurring seed storage protein is genomic DNA.

6. A method according to Claim 2 wherein the DNA sequence is: SEQ ID NO:1

or; a DNA sequence at least 80% homologous thereto.

7. A method according to Claim 3 wherein the DNA sequence is: SEQ ID NO:1

or; a DNA sequence at least 80% homologous thereto.

8. A method according to Claim 2 wherein the DNA sequence is synthesized by the steps of:

a. synthesizing a set of single-stranded partial DNA sequences capable of being assembled in complementary overlapping relationship to provide the complete DNA sequence of the altered protein, each partial sequence having a length of less than about 100 base pairs, each partial sequence having 3′ and 5′ oligonucleotide ends which are complementary to the respective 3′ and 5′ oligonucleotide ends of the partial sequences which are respectively 3′ and 5′ to the partial sequence in the complete sequence of the altered protein; and

b. annealing the partial sequences to produce extended sequences consisting of two or more partial sequences in complementary overlapping relationship;

c. filling nucleotide gaps in the extended sequences to produce double-stranded extended sequences;

d. denaturing the double-stranded extended sequences to produce longer sequences consisting of two or more partial sequences; and

e. repeating steps (b) through (d) until the extended sequence produced by step (c) is the complete DNA sequence of the altered protein.

9. A method of synthesizing a complete DNA sequence comprising the steps of:

a. synthesizing a set of single-stranded partial DNA sequences capable of being assembled in complementary overlapping relationship to provide the complete DNA sequence, each partial sequence having 3′ and 5′ ends which are complementary to the respective 3′ and 5′ ends of the partial sequences which are respectively 3′ and 5′ to the partial sequence in the complete sequence; and

b. annealing the partial sequences to produce extended sequences consisting of two or more partial sequences in complementary overlapping relationship;

c. filling nucleotide gaps in the extended sequences to produce double-stranded extended sequences;

d. denaturing the double-stranded extended sequences to produce longer sequences consisting of two or more partial sequences; and

e. repeating steps (b) through (d) until the extended sequence produced by step c is the complete DNA sequence.

```
     PstI
TAACTGCAG ATG GCA AAC ATT TCT GTG GTT GCT GCT GCA CTA CTG GTC      48
          Met Ala Asn Ile Ser Val Val Ala Ala Ala Leu Leu Val
           1               5                  10

TTG CTG GTG TTG GGT CAT GCC ACT GCA AGC ATC TAC AGG ACA GTT GTG    96
Leu Leu Val Leu Gly His Ala Thr Ala Ser Ile Tyr Arg Thr Val Val
    15                  20                  25

GAG TTT GAA GAG GAT GAT GCC ACC AAC CCA ATA GGT CCT AAG ATG AGG   144
Glu Phe Glu Glu Asp Asp Ala Thr Asn Pro Ile Gly Pro Lys Met Arg
 30                  35                  40                  45

AAA TGC AGA AAG GAG TTC CAG AAG GAA CAA ATG TTG AGA GCT TGC CAA   192
Lys Cys Arg Lys Glu Phe Gln Lys Glu Gln Met Leu Arg Ala Cys Gln
 50                  55                  60                  65

CAA TGG TTG AGG AAA CAA GCT AGA CAA GGA AGA TCT GAT GAA TTT GAC   240
Gln Trp Leu Arg Lys Gln Ala Arg Gln Gly Arg Ser Asp Glu Phe Asp
 70                  75                  80                  85

TTT GAA GAT GAC ATG GAG AAT CCT CAA GGA CCA CAG CAG AGA CCT CCT   288
Phe Glu Asp Asp Met Glu Asn Pro Gln Gly Pro Gln Gln Arg Pro Pro
                 90                  95                 100

CTC CTT CAG AAG TGC TGT GAG CAA CTC AAA CAG ATG CAA TCT CAG TGT   336
Leu Leu Gln Lys Cys Cys Glu Gln Leu Lys Gln Met Gln Ser Gln Cys
            105                 110                 115

GTT TGC CCA ACC CTT AAA GGT GCC AGC AAA GCT GTG AAA CAG GAA GAG   384
Val Cys Pro Thr Leu Lys Gly Ala Ser Lys Ala Val Lys Gln Glu Glu
        120                 125                 130

CAG CAA CAA GGC CAG CAA CAA GGT AAG CAG CAG ATG GTT AGG AAG ATC   432
Gln Gln Gln Gly Gln Gln Gln Gly Lys Gln Gln Met Val Arg Lys Ile
    135                 140                 145

TAT AAG ACT GCC AAA CAC CTT CCT AAA GTC TGT GAC ATT CCA CAG GTT   480
Tyr Lys Thr Ala Lys His Leu Pro Lys Val Cys Asp Ile Pro Gln Val
150                 155                 160                 165
                                                 EcoRI
GAT GTA TGC CCA TTT CAG AAG ACC ATG CCT GGG CCC TCA TAC TAGAATT   529
Asp Val Cys Pro Phe Gln Lys Thr Met Pro Gly Pro Ser Tyr ***
                170                 175

CAAT                                                              533
```

FIGURE 1

TAACTGCAGATGGCAAACATTTCTGGTTGTGCTGCTGCACTACTGGTCTTGCTGGTGTTGGGTCATGCC
GACCACAACCCAGTACGGT

GAAGAGGATGATGCCACCAACCCAATAGGTCCTAAGAT
GACGTTCGTAGATGTCCTGTCAACACCTCAAACTTCTCCTACTACGGTGG

GAGGAAATGCAGAAAGGAG GAGCTTGCCAACAATGGTTGAGGAAACAA
CCTTTACGTCTTTCCTCAAGGTCTTCCTTGTTTACAACTCTCGAACGGTTGTTACC

GCTAGACAAGGAAGATCTGATGAATTTGAC GGACCACAGCAGA
CTAGACTACTTAAACTGAAACTTCTACTGTACCTCTTAGGAGTTCCTGGTGTCGTCT

GACCTCCTCTCCTTCAGAAGTGCTGTGAGCAACTCAAACAGATG
CTGG CTCGTTGAGTTTGTCTACGTTAGAGTCACACAAACGGGTTGGGA

AAGGTGCCAGCAAAGCTGTGAAACAGGAAGAGCAGCAACAAGGCCAGCAACAAGGTAAGCAGCAG
ATTTCCACGGTCGTTTCGAC GTTGTTCCGGTCGTTGTTCCATTCGTCGTCTAC

GACTGCCAAACACCTTCCTAAAGTCTGTGACATTCCACAGGTTGATGTATGCC
CAATCCTTCTAGATATTCTGACGGTTTGTGGAAGG CCAACTACATACGG

CATTTC
GTAAAGTCTTCTGGTACGGACCCGGGAGTATGATCTTAAGTTA

FIGURE 2

Sequential extension from overlap regions of oligonucleotides:

First extension

```
------ - -> Oligo pair 1+/1-
<- - ------
         ------ - -> Oligo pair 2+/2-
         <- - ------
                  ------ - -> Oligo pair 3+/3-
                  <- - ------
                           ------ - -> Oligo pair 4+/4-
                           <- - ------
                                    ------ - -> Oligo pair 5+/5-
                                    <- - ------
                                             ------ - -> Oligo pair 6+/6-
                                             <- - ------
```

Second extension

```
---------- --- ----> Oligo pairs 1 + 2
<---- --- ----------
          ---------- --- ----> Oligo pairs 2 + 3
          <---- --- ----------
                    ---------- --- ----> Oligo pairs 3 + 4
                    <---- --- ----------
                              ---------- --- ----> Oligo pairs 4 + 5
                              <---- --- ----------
                                        ---------- --- ----> Oligo pairs 5 + 6
                                        <---- --- ----------
```

FIGURE 3

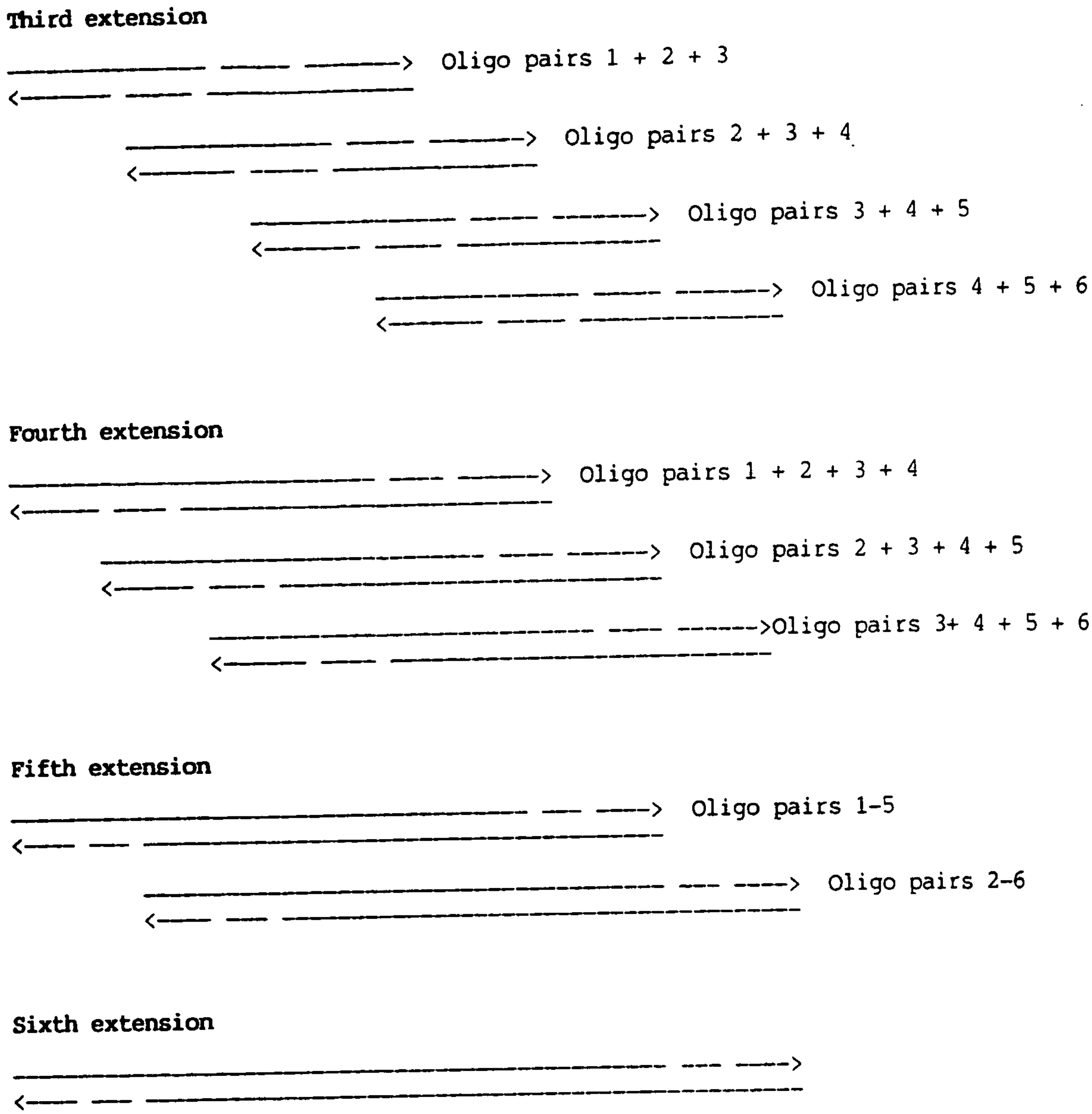
Third extension
Oligo pairs 1 + 2 + 3
Oligo pairs 2 + 3 + 4
Oligo pairs 3 + 4 + 5
Oligo pairs 4 + 5 + 6
Fourth extension
Oligo pairs 1 + 2 + 3 + 4
Oligo pairs 2 + 3 + 4 + 5
Oligo pairs 3+ 4 + 5 + 6
Fifth extension
Oligo pairs 1-5
Oligo pairs 2-6
Sixth extension

FIGURE 4

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgystan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo			SE	Sweden
CH	Switzerland	KR	Republic of Korea	SI	Slovenia
CI	Côte d'Ivoire	KZ	Kazakhstan	SK	Slovakia
CM	Cameroon	LI	Liechtenstein	SN	Senegal
CN	China	LK	Sri Lanka	TD	Chad
CS	Czechoslovakia	LU	Luxembourg	TG	Togo
CZ	Czech Republic	LV	Latvia	TJ	Tajikistan
DE	Germany	MC	Monaco	TT	Trinidad and Tobago
DK	Denmark	MD	Republic of Moldova	UA	Ukraine
ES	Spain	MG	Madagascar	US	United States of America
FI	Finland	ML	Mali	UZ	Uzbekistan
FR	France	MN	Mongolia	VN	Viet Nam
GA	Gabon				